Establishing Correspondences between Attribute Spaces and Complex Concept Spaces Using Meta-PGN Classifier

Krassimira Ivanova[1], Iliya Mitov[1], Peter Stanchev[1,2]
Phillip Ein-Dor[3], Koen Vanhoof[4]

[1]Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Sofia, Bulgaria
{kivanova,imitov}@math.bas.bg
[2]Kettering University, Flint, USA
pstanche@kettering.edu
[3]The Academic College of Tel Aviv-Yafo, Israel
eindor@post.tau.ec.il
[4]Hasselt University, Belgium
koen.vanhoof@uhasselt.be

Abstract. In this paper, we present one approach for extending the learning set of a classification algorithm with additional metadata. It is used as a base for giving appropriate names to found regularities. The analysis of correspondence between connections established in the attribute space and existing links between concepts can be used as a test for creation of an adequate model of the observed world. Meta-PGN classifier is suggested as a possible tool for establishing these connections. Applying this approach in the field of content-based image retrieval of art paintings provides a tool for extracting specific feature combinations, which represent different sides of artists' styles, periods and movements.

Keywords: Multimedia Semantics, Metadata, Data Mining, Pattern Recognition, Classification, Categorization, Content-Based Image Retrieval (CBIR)

1 Introduction

The problem of resolving the gaps between computer analysis and human understanding has a deep philosophical background even in the problem of understanding between humans. Lewis Carroll in "Through the Looking-Glass" gives us an example of absurd use of semantics and pragmatics when Humpty Dumpty talks with Alice: "When I use a word it means just what I choose it to mean – neither more nor less".

At the base of semantics lies the definition of concepts as names and corresponding content. Our life is full with learning concepts (their names and contents) in order to understand each other. There exist many well-suited theories in the area of concept formation based on the attribute models. Very close to this understanding is formal concept analysis [13], that uses the philosophical view of a concept as a unit consist-

ing of two parts – extension (covers all objects belonging to this concept) and intension (comprises all attributes valid for all those objects) [12] and the possibility of using attribute descriptions of the extension of one concept to form its intension definition as well as to uncover the relations between different concepts, like the hierarchical "subconcept-superconcept" and others.

In recent years, when computer systems become part of our interaction, we are faced with the necessity of building an intelligent interface, which includes the use of the concepts in the computer systems in the same way that we use them. However, when the computer finds a typical combination of attributes, which defines something, it does not know how to name it.

It should be mentioned that the problems of semantics and semiotics cannot be reduced only to the interface – they concern expressing an "understanding of the world" by the intelligent system. These problems are well discussed by A. Scherp and R. Jain in their work "Towards an ecosystem for semantics" [10].

In other words, we give the system some description of the real world and expect that it will follow our mental model and will generate appropriate names of concepts that arise. At the same time we expect that the system will explain the decision it has made and will show these parts of the mental model, which it has used to generate the names of the concepts. This interaction with the intelligent system helps us to improve our mental model and to develop it by extending and/or changing any of its parts. Analyzing the answers of the system may show that our mental model does not correspond to the real world. As a result of this process of information interaction with an intelligent system one obtains the possibility of improving the attribute space of some observed area.

The rest of this paper presents a brief explanation of the proposed algorithm in the light of possible implementation in the field of art painting classification. Finally, some conclusions are highlighted.

2 Algorithm Description

We often strive to choose class-section in such way that all values to be subordinated to a common rule, which comes from our sense of order. For instance, if we want to divide the space of art paintings, we usually use such class as "movement" or "artist's name", etc. However, the analysis of colour distribution for different artists [7] shows that there are some artists, for instance Gauguin and Giotto, which use similar colours in all his life and the others, like Velazquez and Rubens, have great disperse of used palettes. In addition, this can be a signal that the work of such authors has to be divided in more short periods. Typical example of this direction is the variety of Picasso' styles.

Below the application of the algorithm for multiple-class categorization (firstly presented in [6]) in the field art-painting image analysis is presented.

2.1 Dataset Construction

The observed set of objects is described by a set of primary measurable features and is classified from a number of viewpoints – a complementary set of classifiers. All this information is given in a form of a table, which consists of two parts – a descriptive part, and a metadata part each consisting of several columns (Figure 1).

The primary features, which participate in the descriptive part in our case, are automatically extracted for given object, derived from the low and intermediate semantic analyses of the paintings. We use different types of attributes. Some of them represent characteristics of some low level visual data. Other attributes are derived as a result of clustering of MPEG-7 descriptors of tiles of paintings. Others are derived from intermediate semantic analysis such as colour harmonies and contrasts [8].

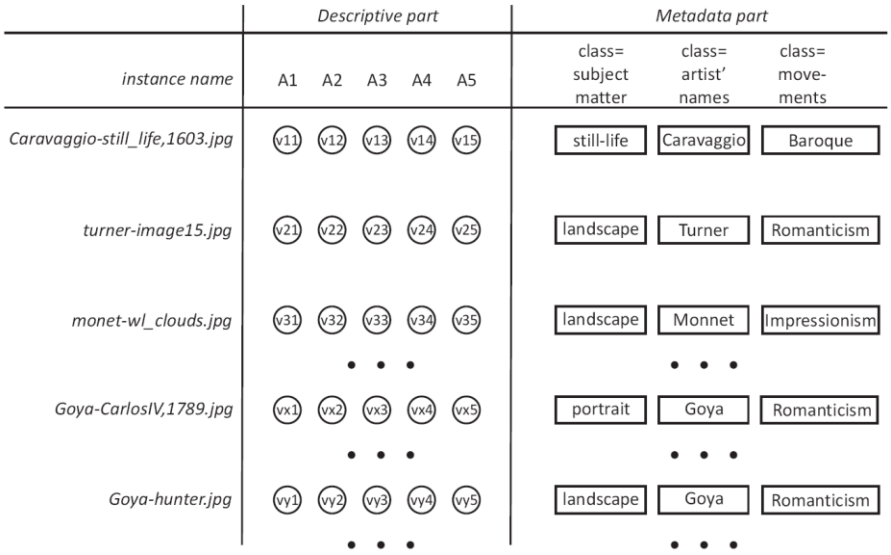

Figure 1. Dataset construction

The metadata part contains classification values of different viewpoints – for example: the subject of the painting (landscape, portrait, scene, still life, etc.), the artist's name (Leonardo, Rubens, Picasso, etc.), the movement (Gothic, Renaissance, Impressionism, etc.), and group of movements (which span movements with common painting techniques), and so on. The metadata can be:

- automatically extracted from the context – for instance in some cases the artist's name is in the file name of the image or in the text of the web-page that presents the image;
- manually annotated – for example the subjects of the paintings can be put in by an expert;
- some of the features can be obtained as a result of secondary processing using already given features and corresponding thesauri – for instance the movement, which the painting belongs to, is determined on the basis of the artist

(sometimes using the year of the paintings) on the one hand, and the previously defined ontology of movements, schools and artists on the other hand.

2.2 Modelling the Interconnections between Objective Categorizations

We address the problem of understanding and modelling the realistic interconnections between an objective categorization given by a set of classifiers and initial object descriptions given by sets of features and native relations of these descriptions.

Some of the concepts can be described in common hierarchical structure and then more abstract concepts inherit properties of their successors. Other concepts may not enter in this hierarchical description as they are connected with different (non-hierarchical) connections with other concepts.

The process of searching for typical combinations, which have to be associated with some concepts, has two sides. From one side this is an easy tool for preliminary analysis of the object features and their combinations and class representatives (from a different point of view), which is interesting for examining the attribute space. This does not eliminate the factor analysis algorithms, but vastly reduces their task by throwing off a part of the variants and proposing the combinations to be further examined. From the other side the processes of knowledge formation can be extended by supplying additional metadata, which are used as a basis for finding appropriate names of defined concepts.

For the purposes of the task in question, the most appropriate algorithms for creating the rules are associative classifiers in the area of association rule mining. The advantage of associative classifiers in this case is the possibility to discover specific attribute combinations that most appropriately describe different class labels. In this case, class labels are not only from one class, but also from different classes.

2.3 Identifying the Concepts

We propose an extension of classical classification methods with the idea of using metadata for automatic concept naming of the regularities found by the system. For implementing the proposed idea we use as a base a part of the already realized classification algorithm in the experimental system PaGaNe (PGN) [9].

The standard classification algorithm PGN uses feature vectors (instances), which consist of the instance name (optional), the class-label of the given instance, as well as a set of values of attributes that characterize the instance. The algorithm finds associative combinations of attribute values that are representative for the corresponding class-label.

The enhanced algorithm meta-PGN uses feature vectors with different structure – all vectors belong to the one generic class, which represents the examined area as a whole. Beside this, the vectors contain additional part of values of metadata domains.

The algorithm contains the following steps:

First step: generating the associative rules – the learning set is processed by the standard classification algorithm PGN in the phase of generating the associative rules. As a result we receive a set of rules, which define specific frequently occurring com-

binations of attributes. Each rule is connected with the instances of the learning set, which were participated in its creation.

Second step: finding metadata values as potential names of the rules – the rules, support for which is more than some threshold given as a parameter in each case, are examined. For each rule, the set of its instances is processed. The second parts of the vectors of these instances are analyzed. For each metadata position the normalized most frequently encountered value is determined. Each of these values is a potential candidate for the name of the concept, which is defined by the typical combination of attributes, contained in the examined rule (Figure 2).

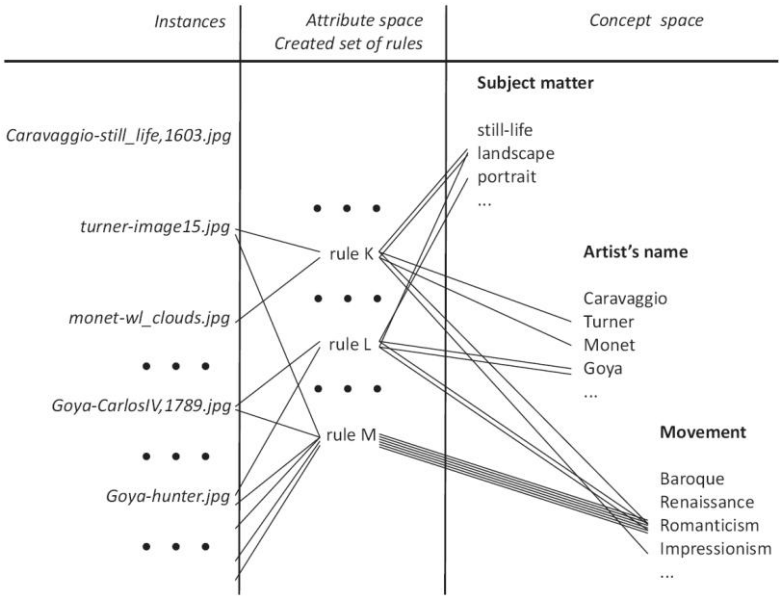

Figure 2. Example of "identifying the concepts" using the support of the rules (landscape – rule K; Goya – rule L; Romanticism – rule M)

Third step: defining the metadata values – all metadata positions are traversed in order to find the rules that correspond to the value of this position. The rules, connected with the examined metadata value, are additionally processed in order to throw out the rules that are supersets of other rules in the group. As a result, for each metadata value (that defines some concept) there are: zero, one, or more corresponding rules. The reason that corresponding rules do not exist usually lies in the fact that: (1) the threshold of this position was too high or (2) chosen primary attributes are not enough to correctly define this concept. If a metadata value is connected with only one rule – we can assume that this is the exact name of this rule. The content of this concept is represented as a conjunction of significant values of attributes, contained in the corresponding rule. There is also a risk that primary attributes do not correctly represent the examined area. However, this is the problem of classification in general.

If the value is connected with more rules – it is represented as a disjunction of these rules.

2.4 Finding Connections between Different Classes of Metadata

The analysis can be continued in order to search for connections between two classes. This realization of meta-PGN algorithm searches for:

- hierarchical dependencies between two classes – if (1) each value in the first class corresponds to some value in the second class; (2) different values in the first class correspond to one value in the second class; (3) there are no different values in the second class that correspond to one value in the first class. In this case the system determines that the first class is hierarchically subordinate to the second class;
- equivalency between two classes – when different values in the first class correspond to different values in the second class and vice versa.

This analysis is focused not to establish the new connections between concepts, but to test the correctness of the chosen attribute space and model of creating the rules in order to reproduce an adequate model of the observed area.

3 Conclusion

Application of some concepts, already known in the pattern recognition area, to solve some novel specific problems of categorization such as discovering the relations between descriptive values and metadata values of the input table in order to use the metadata for automatic concept identification of the found regularities, has been discussed in the paper. One of the advantages of the associative classifiers (respectively associative rule miners) is their ability to discover existence of typical specific combinations of the attribute values.

Our attention in this position paper stopped on the boundary between the processes of clustering that reveal the structure of the attribute space and the processes of categorization that build the bridges between the attribute space and concepts on the way to examining more complex concept spaces. The analysis of correspondence between connections established in the attribute space and existing links between concepts can be used as a test for creation of an adequate model of the observed world. As an example of possible application, the attribute space created on the base of CBIR of paintings and the concept space of art collections is presented.

Acknowledgment. This work is partially financed by the Bulgarian National Science Fund under the joint research project "High Level Semantic Analysis of Bulgarian and Israel Collections" between the Bulgarian Academy of Sciences and the Faculty of Management, Tel Aviv University.

References

1. Agrawal R: Narrowing Down the Semantic Gap between Content and Context Using Multimodal Keywords. PhD thesis, Wayne State University (2009)
2. Castelli, V., Bergman, D. (eds.): Image Databases: Search and Retrieval of Digital Imagery, John Wiley & Sons (2002)
3. Chen, C.-C., Wactlar, H., Wang, J., Kiernan, K.: Digital imagery for significant cultural and historical materials – An emerging research field bridging people, culture, and technologies, Int. J. Digital Libraries 5(4), 275-286 (2005)
4. Croft, W.: What Do People Want from Information Retrieval? (The Top 10 Research Issues for Companies that Use and Sell IR Systems), Centre for Intelligent Information Retrieval Computer Science Department, University of Massachusetts, Amherst (1995)
5. Datta, R.: Semantics and Aesthetic Inference for Image Search: Statistical Learning Approaches, PhD thesis, Pennsylvania State University (2009)
6. Ivanova, Kr., Mitov, I. Markov, Kr., Stanchev, P., Vanhoof, K., Aslanyan, L., Sahakyan, H.: Metric categorization relations based on support system analysis, Proc. of the 7th Int. Conf. "Computer Science and Information Technologies", Yerevan, Armenia, 85-88 (2009)
7. Ivanova, Kr., Stanchev, P., Dimitrov, B.: Analysis of the distributions of color characteristics in art painting images, Serdica J. of Computing, 2(2), 101-126 (2008)
8. Ivanova, Kr., Stanchev, P., Vanhoof, K.: Automatic tagging of art images with color harmonies and contrasts characteristics in art image collections, Int. J on Advances in Software, 3(3&4), 474-484 (2010)
9. Mitov, I., Ivanova, Kr., Markov, Kr., Velychko, V., Vanhoof, K., Stanchev, P.: PaGaNe – A classification machine learning system based on the multidimensional numbered information spaces, Proc. of 4th Int. Conf. "Intelligent Systems and Knowledge Engineering" (ISKE 2009), Hasselt, Belgium, Printed in World Scientific Proceedings Series on Computer Engineering and Information Science, No:2, 279-286 (2009)
10. Scherp, A. Jain, R.: Towards an ecosystem for semantics, Proc. of First ACM Workshop on the Many Faces of Multimedia Semantics (MS'07), 3-11 (2007)
11. Smeulders, A., Worring, M., Santini, S., Gupta, A., Jain, R.: Content based image retrieval at the end of the early years, IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(12), 1349-1380 (2000)
12. Wagner, H.: Begriff, Handbuch Philosophischer Grundbegriffe, München, 191-209 (1973)
13. Wille, R.: Concept lattices and conceptual knowledge systems, Int. J. "Computers and Mathematics with Applications", 23(6-9), 493-515 (1992)